Fundamentals of Machine Learning for Mass Spectrometry Data Analysis

2-Day Short Course

Wout Bittremieux
University of Antwerp

Will Fondrie
Talus Bioscience

Machine Learning for Mass Spectrometry

Course Overview

This course will introduce the fundamentals of common machine learning techniques used for the analysis of mass spectrometry data. The main goal of the course is to promote basic data literacy for people new to machine learning. This will help researchers to perform basic machine learning analyses, know when and how to consult a machine learning expert, and better understand machine learning applications in the scientific literature.

Target Audience

Mass spectrometry researchers who want to obtain a deeper understanding of basic machine learning concepts, algorithms, and common pitfalls.

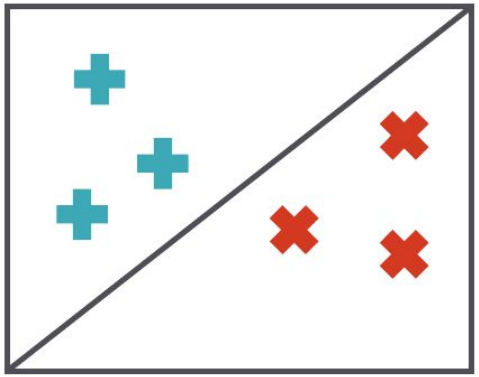
Prerequisites

- Basic knowledge of mass spectrometry data analysis is recommended.
- Basic knowledge of Python programming is recommended.
- No prior machine learning experience is needed.
- Bring a laptop for the example exercises.

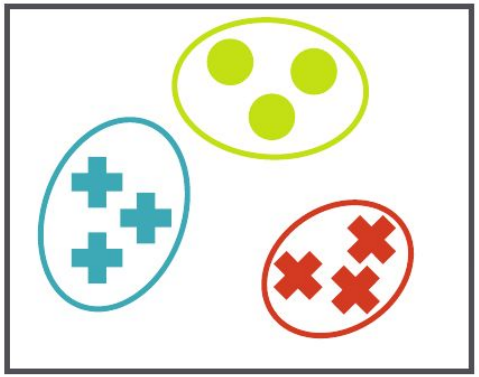
Machine Learning for Mass Spectrometry

Introduction: What is machine learning?

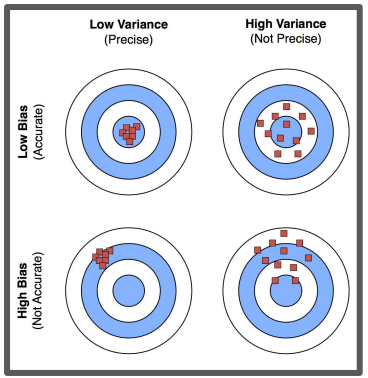
Supervised Learning



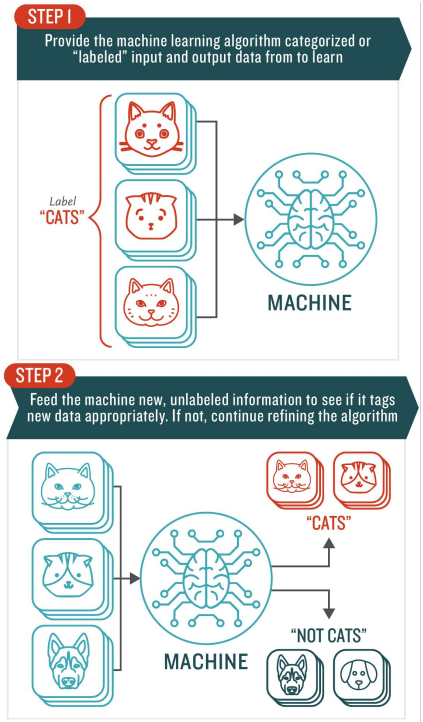
Unsupervised Learning



Model Evaluation



Supervised Learning

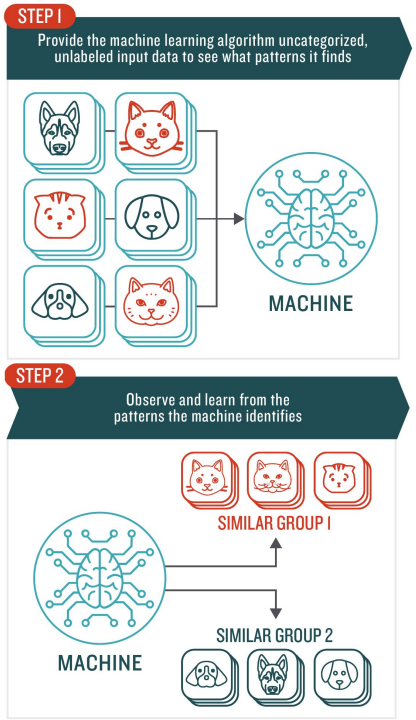


Learn how supervised learning can be used to classify data between different classes.

Introduction to common classification algorithms based on popular MS applications:

- **Logistic regression**: Predict the performance of an MS experiment from quality control metrics.
- **Support vector machine**: Use Percolator to distinguish between true and false peptide-spectrum matches.
- **Decision tree and random forest**: Detect biomarkers from omics features.
- **Neural network**: Predict peptide collisional cross section from timsTOF data

Unsupervised Learning

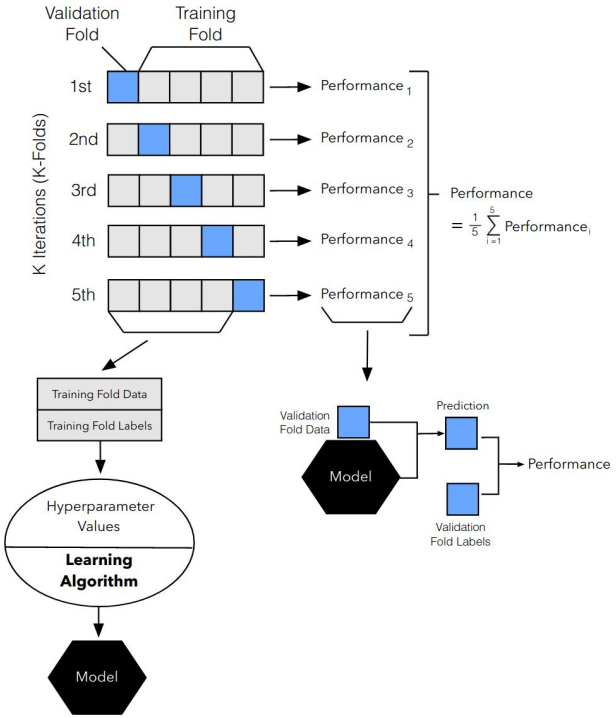


Learn how unsupervised learning algorithms automatically discover patterns in the data.

Introduction to common unsupervised learning techniques based on popular MS applications:

- **Dimensionality reduction**: Visualize high-dimensional data using PCA/PCoA, t-SNE, and UMAP for exploratory data analysis.
- **Clustering**: Condense your data by clustering near-identical MS/MS spectra.

Model Evaluation



Learn best practices when using machine learning. Be aware of potential pitfalls of machine learning and learn to critically assess the scientific literature.

- **Cross-validation**: Train a robust classifier by using a proper cross-validation set-up.
- **Hyperparameter tuning**: Maximize the machine learning performance by finding optimal hyperparameters.
- **Model evaluation**: Assess the performance of a machine learning model using relevant evaluation criteria, such ROC and precision–recall curves.

Additional Course Info

Participants will achieve the following learning goals:

- Recognize when machine learning methods may be beneficial for their research.
- Identify common pitfalls in the application of machine learning methods.
- Gain confidence to provide constructive feedback for applications of machine learning in the manuscripts they review.
- Evaluate the strengths and weaknesses of machine learning approaches presented in the scientific literature.
- Gain familiarity with additional resources to deepen their understanding of machine learning.